SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1064–1071
doi: 10.1007/s11427-014-4747-6

Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks

LI Min[1], LI Qi[1], GANEGODA Gamage Upeksha[1], WANG JianXin[1*], WU FangXiang[1,2] & PAN Yi[1,3]

[1]*School of Information Science and Engineering, Central South University, Changsha 410083, China;*
[2]*College of Engineering, University of Saskatchewan, Saskatoon, SK STN 5A9, Canada;*
[3]*Department of Computer Science, Georgia State University, Atlanta, GA 30302-4110, USA*

Received May 23, 2014; accepted July 15, 2014; published online October 14, 2014

Identification of disease-causing genes among a large number of candidates is a fundamental challenge in human disease studies. However, it is still time-consuming and laborious to determine the real disease-causing genes by biological experiments. With the advances of the high-throughput techniques, a large number of protein-protein interactions have been produced. Therefore, to address this issue, several methods based on protein interaction network have been proposed. In this paper, we propose a shortest path-based algorithm, named SPranker, to prioritize disease-causing genes in protein interaction networks. Considering the fact that diseases with similar phenotypes are generally caused by functionally related genes, we further propose an improved algorithm SPGOranker by integrating the semantic similarity of gene ontology (GO) annotations. SPGOranker not only considers the topological similarity between protein pairs in a protein interaction network but also takes their functional similarity into account. The proposed algorithms SPranker and SPGOranker were applied to 1598 known orphan disease-causing genes from 172 orphan diseases and compared with three state-of-the-art approaches, ICN, VS and RWR. The experimental results show that SPranker and SPGOranker outperform ICN, VS, and RWR for the prioritization of orphan disease-causing genes. Importantly, for the case study of severe combined immunodeficiency, SPranker and SPGOranker predict several novel causal genes.

disease-causing genes, prioritization, gene ontology, protein interaction network, shortest path

Citation: Li M, Li Q, Ganegoda GU, Wang JX, Wu FX, Pan Y. Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks. Sci China Life Sci, 2014, 57: 1064−1071, doi: 10.1007/s11427-014-4747-6

An orphan disease (OD) is the disease that affects only a small percentage of the population, which is defined as a disease that affects fewer than 200000 inhabitants in the USA [1]. There are about 8000 ODs and most of which are genetic, and thus are present throughout the person's entire life, even if symptoms do not immediately appear [2,3]. More important, most ODs affect children at a very early age and about 30% of children with ODs will die before reaching their fifth birthday.

To find disease-causing genes (or genes involved in a biological process), researchers usually use the traditional positional cloning approaches or high-throughput genomic technologies to identify hundreds or even thousands of candidate genes [4]. However, it is too time-consuming and laborious to validate these candidate genes one by one through biological experiments [5,6]. Fortunately, the hu-

*Corresponding author (email: jxwang@csu.edu.cn)

man genome project has been completed and this project has achieved great success, at the same time, the development of high-throughput approaches has provided a large number of protein-protein interactions, which make it possible for us to study life activity at the network level [7–9]. Many network-based methods have been proposed to predict protein functions, identify essential proteins, and detect disease-causing genes and related complexes or pathways [10–21]. It also has been shown that genes associated with the similar phenotypes tend to share the common molecular signatures including similar expression profiles, participation in the same biological processes, pathways, or complexes [22–24]. Moreover, the research of human disease and the experiment of model organisms show that direct and indirect interactions occurred between protein pairs may be responsible for similar disease phenotypes [25–27]. These researches motivate us to predict disease-causing genes by using protein interaction networks.

In recent years, many network-based computational approaches have been proposed for prioritizing candidate disease genes [28–36]. Generally, the prioritization approaches can be grouped into two main categories: (i) global methods which model the information flow in the cell to assess the proximity and connectivity between known disease genes and candidate genes, such as PRIoritizatioN and Complex Elucidation (PRINCE [32]), Random Walk with Restart (RWR [33]), and PageRank with Priors (PRP [34]), and (ii) localized methods which predict new disease candidates by counting the direct interacting genes or computing the shortest paths between known disease genes and candidate genes, such as Interconnectedness (ICN [35]), and Vertex similarity-based frameworks (VS [36]).

In addition, some typical graph partitioning methods and clustering approaches, such as GS [37], MCL [38], VI-Cut [39], IPCA [40], MSCF [41], HC-PIN [42], and their improved algorithms, can also be used to discover candidate disease-related genes.

Although great progress has been made on the network-based methods, it is still a challenging task to identify disease-causing genes based on protein interaction network, for there are still a large number of false-positives or negatives existing in the current available protein-protein interaction data [43]. To reduce the effect of the false positives, different types of biological data related to proteins, such as gene expression profiles [44–48], orthology information [49], gene ontology (GO) annotations, have been used in the identification of protein complexes, discovery of essential proteins, prediction of protein functions, or the detection of disease-causing genes.

Based on an overall analysis of 'guilt-by-association' principle and the effectiveness of local network information, we proposed a localized algorithm, SPranker, to prioritize ODs-causing genes. To reduce the effect of the false positives in protein interaction network, we further integrated GO annotations to improve our proposed algorithm SPranker. The improved algorithm, named SPGOranker, prioritizes the disease-causing candidates by combining GO similarity of the candidate genes with their topological similarity. The proposed algorithms SPranker and SPGOranker were tested on the 172 ODs with at least five known causal genes (from Orphanet database [50]). In our experiments, the leave-one-out cross-validation was used to validate the effectiveness of the proposed algorithms. Based on the leave-one-out cross-validation, the proposed algorithm SPranker was shown to outperform three other state-of-the-art algorithms: RWR [33], ICN [35], and VS [36]. The combination of GO annotations contributed to the prioritization of disease-causing genes and SPGOranker was shown to perform better than SPranker. Moreover, SPranker and SPGOranker were applied to predict potential novel candidate genes by prioritization of the immediate neighbors of known OD genes in the human protein interaction network.

1 Methods

1.1 SPranker: shortest path based algorithm to prioritize disease-causing genes

The common assumption of network-based disease-causing gene prioritization methods is that genes that are physically or functionally close to each other tend to be involved in the same biological pathways and have similar effects on phenotypes, which is known as 'guilt-by-association' principle [51]. Hence, the most important thing is how to measure the similarities between the known disease genes and candidate genes in a protein interaction network. In this paper, we use two different strategies to calculate the similarity between proteins: one is for the connected protein pairs and the other is for the proteins which do not have edges connecting them directly.

To describe the proposed algorithm easily, we first give some necessary definitions. A protein interaction network is described as an undirected graph $G=(V,E,W)$, where V represents the set of proteins, an edge $(v_i,v_j)\in E$ denotes that the protein v_i and the protein v_j connect with each other in the network, and $w(v_i,v_j)\in W$ represents the weight of the edge (v_i,v_j). For an unweighted graph G, $w(v_i,v_j)=1$ if and only if the protein v_i and the protein v_j connect each other in the graph G, otherwise $w(v_i,v_j)=0$. Given a protein $v\in V$, its neighbors are the proteins which interact with it in the network. The set of neighbors of a protein v is marked as N_v.

From the topological view, two proteins are considered to be similar if they have more common immediate neighbors in the network. Hence, for a pair of proteins (v_i and v_j) which connect each other in the protein interaction network, we calculate its similarity $Sim^l(v_i, v_j)$ by using the following formula:

$$Sim^t\left(v_i,v_j\right)=\frac{2*w\left(v_i,v_j\right)+\sum_{v_k\in\left(N_{v_i}\cap N_{v_j}\right)}w\left(v_i,v_k\right)*w\left(v_j,v_k\right)}{\sqrt{\sum_{v_k\in N_{v_i}}w\left(v_i,v_k\right)^2}\times\sqrt{\sum_{v_k\in N_{v_j}}w\left(v_j,v_k\right)^2}}, \quad (1)$$

where $w(v_i, v_j)$=1 if the protein v_i and the protein v_j connect each other in the network.

For a pair of proteins (v_i and v_j) which do not have any edges connecting them in the graph G, their similarity, marked as $Sim^{t*}(v_i, v_j)$, is calculated by the following formula:

$$Sim^{t*}\left(v_i,v_j\right)=Sim^t\left(v_i,u_1\right)*\prod_{k=1}^{n-1}Sim^t\left(u_k,u_{k+1}\right)*Sim^t\left(u_n,v_j\right), \quad (2)$$

where u_k denotes the intermediate node on the shortest path from the protein v_i to the protein v_j. If there exist more than one shortest path for a pair of nodes v_i and v_j, all the shortest paths will be kept and the corresponding $Sim^{t*}(v_i, v_j)$ will be calculated and the largest $Sim^{t*}(v_i, v_j)$ will be kept. A schematic drawing of a shortest path from v_i to v_j with n+1 hops is shown in Figure 1.

Generally, the shortest path from a given protein v_i to a target protein v_j was calculated by computing its hops from v_i to v_j [36]. However, a path with the least hops may not be the true path. For example, as shown in Figure 2, there are three paths from the protein v_i to the protein v_j: path 1 (v_i; v_3; v_j) with two hops, path 2 (v_i; v_1; v_2; v_j) with three hops, and path 3 (v_i; v_4; v_5; v_6; v_j) with four hops. The weight in the figure denotes the reliability between two proteins. It will be path 1 if the shortest path is selected by hops. However, path 2 will be selected if we consider the reliability of the edges on the path. It is difficult to be reachable from v_i to v_j if path 1 with the unreliable edge (whose weight is 0.01 and which may be the false positive edge and not there actually) is selected.

Hence, in our approach, for a pair of proteins we try to find the most reliable path between them. That means we must find the path with the highest continued product of edge weight by considering the "bucket effect". To avoid enumeration of all the possible paths between v_i and v_j, we change the problem for calculating the continued product of edge weight for each path into the problem for finding a shortest path between v_i and v_j.

To calculate the shortest path between a pair of two proteins, we use $1/Sim^t(v_i, v_j)$ to describe the distance between two connected proteins v_i and v_j and reweight the edge (v_i, v_j) in the graph G. The shortest path is calculated by using the Dijkstra's algorithm [52].

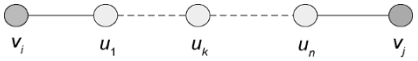

Figure 1 Schematic drawing of a shortest path from a protein v_i to a protein v_j with n+1 hops.

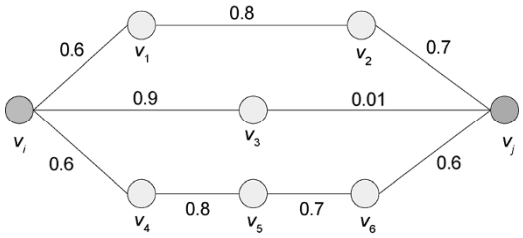

Figure 2 An example of shortest path calculation with multiple paths from a protein v_i to a protein v_j.

Take Figure 2 for example, the lengths of path 1, path 2 and path 3 are 101.11 (1/0.9+1/0.01), 4.35 (1/0.6+1/0.8+1/0.7), and 6.01(1/0.6+1/0.8+1/0.7+1/0.6), respectively. Hence, the shortest path between v_i and v_j is path 2 and it is the most reliable path among the three paths between v_i and v_j in Figure 2. Hence, the similarity of v_i and v_j is calculated by using path 2 not path 1 with the least hops.

For a given disease d, let S_d denote the set of its known disease-causing genes. Then, for a new candidate gene v_i, the probability that it is also such a disease-causing gene is evaluated by the sum of similarities between it and the known disease-causing genes in S_d, as shown in eq. (3):

$$score_{v_i}=\sum_{v_j\in S_d}Sim^t(v_i,v_j). \quad (3)$$

All the candidate genes are then ranked based on these scores. The complexity of calculating a candidate gene's probability to be a disease-causing gene depends on the number of known disease-causing genes in S_d and the complexity of computing the shortest paths.

1.2 SPGOranker: integration of GO annotations into SPranker

It has been shown that similarities among disorders imply involvement of functionally related gene products, which is generally summarized as "phenotypic overlap implies genetic overlap". Recent studies also showed that diseases with similar phenotypes often involve common molecular mechanisms and the functions of their corresponding disease-causing genes are generally similar.

Franke's work [53] shows that GO annotations are the most effective resources for the identification of disease-causing genes. Many other researchers also explored the relationship between the GO annotations and disease-causing genes and predicted new candidates by using their functional similarity to the known causal genes, such as G2D [54], POCUS [55], FP [56] and GFFST [57].

The Gene Ontology project [58] provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data. The GO ontology is with a hierarchical structure and is generally described as a directed acyclic graph (DAG). There have been many com-

putational methods for calculating the similarity between GO terms [59,60]. In this paper, we used one of our previous methods in [59] to calculate the GO term similarity by taking into account the hierarchical organization of functional annotations. For a given pair of proteins v_i and v_j, their functional similarity $Sim^f(v_i, v_j)$ is defined as the maximum similarity of their corresponding GO terms, as shown in eq. (4):

$$Sim^f\left(v_i, v_j\right) = \max_{c_1 \in T_i, c_2 \in T_j} \left(goSim\left(c_1, c_2\right)\right), \quad (4)$$

where T_i and T_j are the corresponding sets of GO terms for the protein v_i and the protein v_j, respectively. The functional similarity, marked as $goSim(c_1,c_2)$, between the term $c_1 \in T_i$ and term $c_2 \in T_j$ is calculated by using the method in [59].

For each species, there are three types of annotations: biological processes, cellular components and molecular functions. In this paper, the molecular functions are used to calculate the functional similarity between proteins and an improved algorithm SPGOranker is proposed to prioritize disease-causing genes by integrating the functional similarity into SPranker. A parameter α is used to integrate the topological similarity and the functional similarity, as shown in eq. (5).

$$Sim\left(v_i, v_j\right) = a * Sim^t\left(v_i, v_j\right) + \left(1-a\right) * Sim^f\left(v_i, v_j\right). \quad (5)$$

The rest steps of SPGOranker are the same as those of SPranker. When α=1, only the topological similarity is considered and SPGOranker will degenerate into SPranker.

2 Results and discussion

2.1 Experimental data sources

The human protein-protein interactions were downloaded from release 9 of the Human Protein Reference Database [61] with both redundant interactions and self-loops removed. Finally, a protein interaction network with 9763 genes and 37060 interactions was obtained and the unweighted network was used in our experiment.

The ODs and causal gene information were downloaded from Orphanet [50]. Thereafter, we merged some of the OD sub-types of a single disease based on the disorder names described in [30,33]. In our work, we selected 172 ODs that have at least five causal genes. The 172 ODs contain 1598 genes in total and 1063 OD-causing genes were found in the protein interaction network used in this paper.

2.2 Cross-validation analysis

To evaluate the performance of the proposed algorithms SPranker and SPGOranker for prioritizing disease-causing genes, we compared them with three state-of-the-art methods: RWR [33], ICN [35], and VS [36]. A leave-one-out cross-validation procedure was used to carry out the evaluation. In each cross validation trial, one causal gene with an OD ("target gene") from the data was removed, and each algorithm was evaluated by its success in assigning the rank to the "target gene". For each validation trial, one seed gene ("target gene") from 172 ODs was removed and mixed with 99 genes selected randomly from the protein interaction network to form a set of 100 candidate genes. The remaining seed genes are treated as the training set.

SPranker, SPGOranker, ICN, VS, and RWR were applied to the test set for prioritizing OD-causing genes. In this paper, the default value of α in SPGOranker is set to be 0.8. During each execution, the rank of the "target gene" was marked. The performance of each algorithm is evaluated in terms of success rate with the respective rank cut-off (k). If the "target gene" is ranked among the top k in a particular validation trial, it is considered as a "success". As what has been done in [36], we also used k ranging from 1 to 30 in this paper. Validation trials are repeated until all the seed genes have been used as the target genes and their ranks are obtained. The "success rate" is defined as the ratio of successful validation trails and the total validation trails for all the existing OD genes from 172 ODs [36].

The results of the five algorithms SPranker, SPGOranker, ICN, VS, and RWR on the success rate with k ranging from 1 to 30 are shown in Figure 3. When k=1, out of 1063 cases 407 are achieved successfully by SPGOranker. SPranker and VS also achieve similar performance with a success rate of 37.8% (402/1063) and 37.1% (394/1063). The success rates of RWR and ICN are 35.9% (381/1063) and 32.7% (348/1063), respectively. From Figure 3 we can see that SPGOranker performs best consistently with k ranging from 1 to 30, and SPranker also achieves a better result compared with ICN, VS, and RWR. These results imply the effectiveness of SPranker for the prioritization of OD-causing genes and the integration of GO annotations contributes to the improvement of detecting true OD-causing genes.

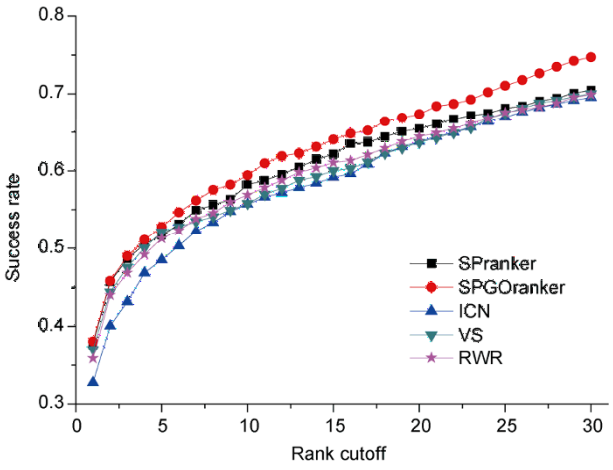

Figure 3 (color online) Comparison of five prioritization algorithms, SPranker, SPGOranker, ICN, VS, and RWR, on the success rate with k ranging from 1 to 30.

To further investigate why and how SPranker and SPGOranker prioritize the disease-causing genes effectively, we analyze the intersection of disease-causing genes identified by the algorithms SPranker, SPGOranker, ICN, VS and RWR with k=1. The analysis results are shown in Table 1.

As shown in Table 1, there are 323 disease-causing genes both identified by SPranker and by SPGOranker, which is about 80.3% of the true predictions by SPranker and by SPGOranker. Out of 394 disease-causing genes identified by VS, 309 genes (about 78.4%) are also discovered by SPGOranker and SPGOranker predicts 96 different true disease-causing genes. ICN identifies 348 disease-causing genes and about 77.0% are covered by SPGOranker. One hundred and thirty-seven different true disease-causing genes are predicted by SPGOranker, which is about 1.7 times that identified by ICN. Out of 381 disease-causing genes predicted by RWR, 288 genes (about 75.6%) are also discovered by SPGOranker and SPGOranker predicts 117 different true disease-causing genes.

2.3 Analysis of the effect of parameter α

In the above analysis, α=0.8 is used in SPGOranker. To analyze the effect of parameter α on the results, we change the value of α from 0.0 to 1.0 with 0.1 increments. When α=0.0, only GO information is used to calculate the similarity between proteins. The analysis results are shown in Figure 4.

Table 1 Contingency table of disease-causing genes identified by the algorithms SPranker, SPGOranker, ICN, VS and RWR with k=1

	SPranker	SPGOranker	VS	ICN	RWR
SPranker	402	323	309	268	287
SPGOranker	323	405	309	268	288
VS	309	309	394	264	285
ICN	268	268	264	348	257
RWR	287	288	285	257	381

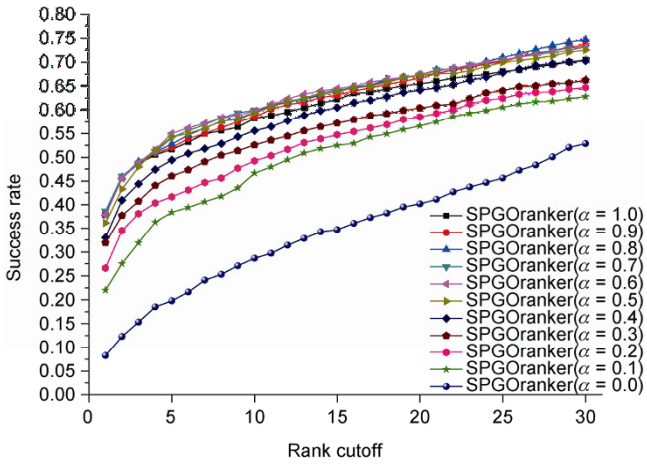

Figure 4 (color online) Cross validation results with different values of the parameter α.

From Figure 4 we can see that with the increase of rank cutoff k, the success rates also increase. For the same rank cutoff k, the success rates of SPGOranker also increase with the increase of α up to 0.6. For all the rank cutoff k=5, 10, 15, 20, 25, 30, similar results are obtained by SPGOranker with α ranging from 0.6 to 0.8. There is a small decline for the success rate when α>0.8 is used. When α=1, a more clear decline for the success rate can be seen from Figure 4 as SPGOranker has degenerated into SPranker. The analysis of parameter α's effect shows that the topological similarity and the functional similarity have their own special contributions to the prioritization of disease-causing genes.

2.4 Predicting novel causing genes of 10 ODs

The proposed algorithms SPranker and SPGOranker were applied on 172 ODs to predict potential novel OD-causing genes. We select ten ODs which are associated with more than 10 known valid causing genes and have known protein-protein interactions for all of their causing genes. The known OD causing genes are considered as seeds and the immediate neighbors of the known causing genes in the protein interaction network are considered as candidate disease genes. Taken the severe combined immunodeficiency (SCID) as case study, the top 10 candidate genes predicted by SPranker, SPGOranker, ICN, VS, and RWR are shown in Table 2.

SCID is a genetic disorder characterized by the absence of functional T-lymphocytes. It is the most severe form of primary immunodeficiency. As shown in Table 2, the first predictions of SPranker and SPGOranker are both "ZAP70 (zeta-chain-associated protein 70 kD)", which is also identified by VS as the top 1. Though ZAP70 is not collected into Orphanet [50], recent research has shown that ZAP70 deficiency is an autosomal recessive form of severe combined immune deficiency (SCID). Moreover, Pagon's work [62] also shows that ZAP70 is signaling of abnormal T cell receptor, which is caused by cell-mediated immunodeficiency.

A new candidate gene "JAK1 (Janus kinase 1)" is identified as the second one both by SPranker and SPGOranker,

Table 2 Top 10 predictions of severe combined immunodeficiency

Rank	SPranker	SPGOranker	ICN	VS	RWR
1	ZAP70	ZAP70	1IL2RB	ZAP70	STIM2
2	JAK1	JAK1	TRAT1	JAK1	IPO5
3	IL2RB	STAT5A	JAK1	STAT5A	TRPC6
4	STAT5A	IL2RB	ZAP70	TRB@	TRPC3
5	CD3G	PTPN6	PTPN22	CD3G	XRCC4
6	TRB@	CD3G	STAT5A	PTPN6	LIG4
7	PTPN6	TRB@	PTPN6	IL2RB	ACTB
8	TRAT1	PTPN22	SHC1	PTPN22	GRB2
9	TRA@	TRAT1	SYK	TRA@	THBS1
10	TSLP	IL7	PIK3R1	TRAT1	FYN

Table 3 Ten ODs and their top 5 predictions[a)]

Disease name	KN	Method	Top 5 predictions
Retinitis pigmentosa	35	SPranker	RDH5, DHX30, SLC24A1, PRPH, FIZ1
		SPGOranker	RDH5, DHX30, SLC24A1, FIZ1, PRPH
Microdeletion syndrome	21	SPranker	FOXP4, FOXP1, SLC7A8, MEF2D, NKX2-5
		SPGOranker	FOXP4, FOXP1, SLC7A8, MEF2D, NKX2-5
Cone rod dystrophy	17	SPranker	CNGB1, CABP4, GUCA2B, NPHP4, ROM1
		SPGOranker	CNGB1, CABP4, NPHP4, GUCA2B, ROM1
Severe combined immunodeficiency	17	SPranker	ZAP70, JAK1, IL2RB, STAT5A, CD3G
		SPGOranker	ZAP70, JAK1, STAT5A, IL2RB, PTPN6
Fanconi anemia	15	SPranker	HES1, XRCC3, SAMD3, CYP19A1, RAD51
		SPGOranker	HES1, XRCC3, RAD51, USP1, CYP19A1
Zellweger syndrome	14	SPranker	PEX7, ABCD1, ABCD2, ABCD3, PXMP4
		SPGOranker	ABCD1, ABCD2, PEX7, ABCD3, PEX11A
Neonatal adrenoleukodystrophy	12	SPranker	PEX7, ABCD1, ABCD2, ABCD3, PXMP4
		SPGOranker	ABCD1, ABCD2, PEX7, ABCD3, PXMP4
Infantile Refsum disease	12	SPranker	PEX7, ABCD1, ABCD2, ABCD3, PXMP4
		SPGOranker	ABCD1, ABCD2, PEX7, ABCD3, PXMP4
Papillary or follicular thyroid carcinoma	11	SPranker	OCRL, TRIM28, RNF14, ZNF10, SHC1
		SPGOranker	OCRL, TRIM28, RNF14, SHC1, GAB1
Romano-Ward syndrome	11	SPranker	KCNE4, ALG10B, NDUFS6, ALG10, KCNJ3
		SPGOranker	KCNE4, ALG10B, KCNJ3, NDUFS6, ALG10

a) KN means the number of the known disease-causing genes in the corresponding OD.

and "IL2RB" is listed as the third one by SPranker and the fourth one by SPGOranker. Russell's study [63] shows that gene JAK1 and gene IL-2Rβ&γ interact with each other, and Ç chain mutations lead to X-linked severe combined immunodeficiency. In addition, "TSLP (thymic stromal lymphocytes)" was listed as the 10th one by SPranker, but did not appear at any top 10 list by any other four algorithms. It has been shown that TSLP plays a role in controlling innate and adaptive immune responses [64].

The top 5 predictions for each OD by SPranker and SPGOranker are shown in Table 3.

3 Conclusion

In this study, we proposed two algorithms SPranker and SPGOranker to prioritize disease-causing genes in protein interaction networks. SPranker is a simple localized algorithm which only uses a protein interaction network. SPGOranker prioritizes candidates by considering both topological similarity and functional similarity between the predicted candidates and the known disease-causing genes. The proposed algorithms SPranker and SPGOranker were applied to 1598 known orphan disease-causing genes (ODGs) from 172 orphan diseases (ODs) and compared with three state-of-the-art approaches, ICN, VS and RWR. The experimental results show that SPranker and SPGOranker outperform ICN, VS, and RWR for the prioritization of orphan disease-causing genes. Importantly, the top predictions of SPranker and SPGOranker for the severe combined immunodeficiency match the known literature, providing further investigation of several novel causal relationships.

It is important to note that the performance of the proposed algorithm SPranker will rely on the quality of the protein interaction network though the only usage of network makes it simple and convenient to be used. For the algorithm SPGOranker, the combination of GO annotations improves its performance for the prioritization of disease-causing genes, but takes extra time to compute the GO term similarity and makes it more complex. Of course, the extra time costs are worthwhile. Moreover, only the molecular functions are considered in this paper. In our future work, the biological processes, cellular components and other biological information will be further considered.

This work was supported in part by the National Natural Science Foundation of China (61370024, 61428209, 61232001) and Program for New Century Excellent Talents in University (NCET-12-0547).

1 Dear JW, Lilitkarntakul P, Webb DJ. Are rare diseases still orphans or happily adopted? The challenges of developing and using orphan medicinal products. British J Clin Pharmacol, 2006, 62: 264–271
2 Schieppati AHJ, Daina E, Aperia A. Why rare diseases are an important medical and social issue. Lancet, 2008, 371: 2039–2041
3 Stolk P, Willemen MJC, Leufkens HGM. Rare essentials: drugs for rare diseases as essential medicines. Bull World Health Org, 2006, 84: 745–751
4 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet, 2003, 33: 228–237
5 Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science, 2002, 298: 2345–2349
6 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for

complex traits: consensus, uncertainty and challenges. Nat Rev Genet, 2008, 9: 356–369
7 Wang J, Li M, Deng Y, Pan Y. Recent advances in clustering methods for protein interaction networks. BMC Genomics, 2010, 11: S10
8 Li M, Wu X, Wang J, Pan Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. BMC Bioinformatics, 2012, 13: 109
9 Zhao B, Wang J, Li M, Wu FX. Detecting protein complexes based on uncertain graph model. IEEE/ACM Trans Comput Biol Bioinform, 2014, doi:10.1109/TCBB.2013.2297915
10 Zhong J, Wang J, Peng W, Zhang Z, Pan Y. Prediction of essential proteins based on gene expression programming. BMC Genomics, 2013, 14: 1–8
11 Wang J, Peng W, Wu FX. Computational approaches to predicting essential proteins: a survey. Proteomics Clin Appl, 2013, 7: 181–192
12 Peng W, Wang J, Cai J, Chen L, Li M, Wu FX. Improving protein function prediction using domain and protein complexes in PPI networks. BMC Syst Biol, 2014, 8: 35
13 Wang J, Ren J, Li M, Wu FX. Identification of hierarchical and overlapping functional modules in PPI networks. IEEE Trans Nano-Biosci, 2012, 11: 386–393
14 Wang J, Liu B, Li M and Pan Y. Identifying protein complexes from interaction networks based on clique percolation and distance restraction. BMC Genomics, 11: S10
15 Li M, Wang J, Chen J, Cai Z, Chen G. Identifying the overlapping complexes in protein interaction networks. Int J Data Min Bioinform, 2010, 4: 91–108
16 Peng W, Wang J, Cheng Y, Lu Y, Wu FX, Pan Y. UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform, 2014, doi: 10.1109/TCBB.2014.2338317
17 Zhao B, Wang J, Li M, Wu FX, Pan Y. Prediction of essential proteins based on overlapping essential modules. IEEE Trans NanoBiosci, 2014, doi: 10.1109/TNB.2014.2337912
18 Li M, Wang J, Wang H, Pan Y. Identification of essential proteins from weighted protein interaction networks. J Bioinform Comput Biol, 2013, 11: 1341002
19 Wang J, Li M, Wang H, Pan Y. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 1070–1080
20 Li M, Zhang H, Wang J, Pan Y. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. BMC Syst Biol, 2012, 6: 15
21 Li M, Wang J, Chen X, Wang H, Pan Y. A local average connectivity-based method for identifying essential proteins from the network level. Comput Biol Chem, 2011, 35: 143–150
22 Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011, 12: 56–68
23 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA, 2007, 104: 8685–8690
24 Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci USA, 2008, 105: 4323–4328
25 Oti M, Brunner HG. The modular nature of genetic diseases. Clin Genet, 2007, 71: 1–11
26 Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyansky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. The genomic landscapes of human breast and colorectal cancers. Science, 2007, 318: 1108–1113
27 Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási AL, Vidal M, Zoghbi HY. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell, 2006, 125: 801–814
28 Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics, 2010, 26: 1057–1063
29 Ganegoda GU, Wang J, Wu FX, Li M. Prediction of disease genes using tissue-specified gene-gene network. BMC Syst Biol, 2014, 8(Suppl 3): S3
30 Wang J, Chen G, Li M, Pan Y. Integration of breast cancer gene signatures based on graph centrality. BMC Syst Biol, 2011, 5: S10
31 Chen B, Wang J, Li M, Wu FX. Identifying disease causing genes by integrating multiple data sources. BMC Med Genom, 2014, 7(Suppl 2): S2
32 Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 2010, 6: e1000641
33 Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 2008, 82: 949–958
34 Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 2009, 10: 73
35 Hsu CL, Huang YH, Hsu CT, Yang UC. Prioritizing disease candidate genes by a gene interconnectedness-based approach. BMC Genomics, 2011, 12: S25
36 Zhu C, Kushwaha A, Berman K, Jegga AG. A vertex similarity-based framework to discover and rank orphan disease-related genes. BMC Syst Biol, 2012, 6: S8
37 Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008. 419–432
38 van Dongen S. Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl, 2008, 30: 121–141
39 Navlakha S, White J, Nagarajan N, Pop M, Kingsford C. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. J Computat Biol, 2010, 17: 503–516
40 Li M, Chen J, Wang J, Hu B, Chen G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinformatics, 2008, 9: 398
41 Ding X, Wang W, Peng X, Wang J. Mining protein complexes from PPI networks using the minimum vertex cut. Tsinghua Sci Technol, 2012, 17: 674–681
42 Wang J, Li M, Chen J, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Trans Computat Biol Bioinform, 2011, 8: 607–620
43 Montanez G, Cho YR. Predicting false positives of protein-protein interaction data by semantic similarity measures. Curr Bioinform, 2013, 8: 339–346
44 Li M, Zheng R, Zhang H, Wang J, Pan Y. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. Methods, 2014, 67: 325–333
45 Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinform, 2014, 11: 407–418
46 Wang J, Peng X, Peng W, Wu FX. Dynamic protein interaction network construction and applications. Proteomics, 2014, 8: 338–352
47 Wang J, Peng X, Li M, Pan Y. Construction and application of dynamic protein interaction network based on time course gene expression data. Proteomics, 2013, 13: 301–312
48 Tang X, Feng Q, Wang J, He Y, Pan Y. Clustering based on multiple biological information: approach for predicting protein complexes. IET Syst Biol, 2013, 7: 223–230
49 Peng W, Wang J, Wang W, Liu Q, Wu FX, Pan Y. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. BMC Syst Biol, 2012, 6: 87
50 Aymé S. Orphanet, an information site on rare diseases. Soins; la revue de référence infirmière, 2003, 672: 46
51 Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of guilt-by-association within gene coexpression net-

works. BMC Bioinformatics, 2005, 6: 227
52 Dijkstra EW. A note on two problems in connexion with graphs. Numerische Mathematik, 1959, 1: 269–271
53 Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet, 2006, 78: 1011–1025
54 Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet, 2002, 31: 316–319
55 Turner FS, Clutterbuck DR, Semple CAM. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol, 2003, 4: R75–R75
56 Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics, 2002, 18: S110–115
57 Zhang P, Zhang J, Sheng H, Russo JJ, Osborne B, Buetow K. Gene functional similarity search tool (GFSST). BMC Bioinformatics, 2006, 7: 135
58 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25–29
59 Li M, Wu X, Pan Y, Wang J. hF-measure: a new measurement for evaluating clusters in protein-protein interaction networks. Proteomics, 2013, 13: 291–300
60 Wang J, Dai L, Li M. GO semantic similarity-based false positive reduction of protein-protein interactions. In: IEEE International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 2009. 211–214
61 Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, Niranjan V, Harsha HC, Talreja N, Vrushabendra BM, Ramya MA, Yatish AJ, Joy M, Shivashankar HN, Kavitha MP, Menezes M, Choudhury DR, Ghosh N, Saravana R, Chandran S, Mohan S, Jonnalagadda CK, Prasad CK, Kumar-Sinha C, Deshpande KS, Pandey A. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 2004, 32: D497–501
62 Ikeda MD, Larkin A. *ZAP70*-related severe combined immunodeficiency. In: Pagon RA, Adam MP, Ardinger HH, Bird TD, Dolan CR, Fong CT, Smith RJH, Stephens K, eds. SourceGeneReviews®. Seattle: University of Washington, Seattle, 2009
63 Russell SM, Johnston JA, Noguchi M, Kawamura M, Bacon CM, Friedmann M, Berg M, McVicar DW, Witthuhn BA, Silvennoinen O. Interaction of IL-2R beta and gamma c chains with Jak1 and Jak3: implications for XSCID and XCID. Science, 1994, 266: 1042–1045
64 Sebastian K, Borowski A, Kuepper M, Friedrich K. Signal transduction around thymic stromal lymphopoietin (TSLP) in atopic asthma. Cell Commun Signal, 2008, 6: 5